2010

Using in-air acoustic vector sensors for tracking moving speakers

Muawiyath Shujau
University of Wollongong, mshujau@uow.edu.au

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

I. Burnett
Faculty of Informatics, University of Wollongong, ianb@uow.edu.au

Using in-air acoustic vector sensors for tracking moving speakers

Abstract

This paper investigates the use of an Acoustic Vector Sensor (AVS) for tracking a moving speaker in real time through estimation of the Direction of Arrival (DOA). This estimation is obtained using the MUltiple SIgnal Classification (MUSIC) [1] algorithm applied on a time-frame basis. The performance of the AVS is compared with a SoundField Microphone which has similar polar responses to the AVS using time-frames ranging from 20 ms to 1 s. Results show that for 20 ms frames, the AVS is capable of estimating the DOA for both mono-tone and speech signals, which are both stationary and moving, with an accuracy of approximately 1.6^0 and less than 5^0 in azimuth, for stationary and moving speech sources, respectively. The results also show that the DOA estimates using the SoundField microphone are significantly less accurate than those obtained from the AVS. Furthermore, the results suggest that for estimating the DOA for speech sources, a Voice Activity Detector (VAD) is critical to ensure accurate azimuth estimation.

Disciplines

Physical Sciences and Mathematics

Publication Details

M. Shujau, C. H. Ritz & I. S. Burnett, "Using in-air acoustic vector sensors for tracking moving speakers," in International Conference on Signal Processing and Communication Systems, 2010, pp. 1-5.

USING IN-AIR ACOUSTIC VECTOR SENSORS FOR TRACKING MOVING SPEAKERS

M. Shujau, C. H. Ritz
School of Electrical, Computer, and Telecommunications Engineering
University of Wollongong, Wollongong NSW Australia
[ms970, critz]@uow.edu.au

I. S. Burnett
School of Electrical and Computer Engineering
RMIT University, Melbourne, VIC, Australia
ian.burnett@rmit.edu.au

Abstract— This paper investigates the use of an Acoustic Vector Sensor (AVS) for tracking a moving speaker in real time through estimation of the Direction of Arrival (DOA). This estimation is obtained using the MUltiple SIgnal Classification (MUSIC) [1] algorithm applied on a time-frame basis. The performance of the AVS is compared with a SoundField Microphone which has similar polar responses to the AVS using time-frames ranging from 20 ms to 1 s. Results show that for 20 ms frames, the AVS is capable of estimating the DOA for both mono-tone and speech signals, which are both stationary and moving, with an accuracy of approximately 1.6^0 and less than 5^0 in azimuth, for stationary and moving speech sources, respectively. The results also show that the DOA estimates using the SoundField microphone are significantly less accurate than those obtained from the AVS. Furthermore, the results suggest that for estimating the DOA for speech sources, a Voice Activity Detector (VAD) is critical to ensure accurate azimuth estimation.

Index Terms: *Microphone arrays, Vector Sensors, Direction of Arrival Estimate*

I. INTRODUCTION

Direction of Arrival (DOA) estimation is important for applications such as video tele-conferencing for automatic camera steering, 3D sound field reproduction and in some military applications [2, 3]. In [4], the use of an Acoustic Velocity Sensor (AVS) for DOA estimation was investigated. The AVS of [4] is used in this study consists of three orthogonally mounted acoustic particle velocity sensors placed orthogonally in the X, Y and Z directions and one omni-directional acoustic pressure sensor occupying a volume of $1cm^3$ and is shown in Fig. 1. A key advantage of the AVS over other microphone arrays is its ability to capture the directional components of sound sources using a very compact array. Accurate DOA estimation is also important for beamforming applied to speech recordings as described in [5].

In [4], the AVS design was considered with a solution presented that resulted in the AVS being capable of producing accurate DOA estimates of mono-tone sources with errors of less than 2^0. While in [4] the results presented were for anechoic conditions with stationary target sources, this work extends to include DOA estimates for speech sources and moving target sources in reverberant environments. The only microphone array that closely resembles the AVS in terms of how the signals are captured is the SoundField Microphone which has four cardioid pressure sensors arranged in a tetrahedron configuration. Unlike the AVS, the SoundField produces the X, Y and Z directional components by combining the four capsule signals. Here, results are compared for DOA estimation using both the AVS and SoundField microphones.

Most of the work done on DOA estimation and speaker tracking is based on the Time Delay Estimates (TDE) or Time Difference of Arrival (TDOA) with non co-incidental microphone arrays. In [6] six pairs of four microphones are used to track and find the DOA estimates using non-linear particle filtering. In [7] 3 microphones are positioned in a straight line to form a microphone array with known geometry and in [7] 3 sound field microphones are arranged in a row and using the X and W components only source localization is achieved. In [8] binaural microphones are used to track multiple speakers in a cocktail party situation.

In reverberant environments, these TDE based approaches are less accurate due to sound reflections. In contrast, since microphones are co-located, the AVS does not relay on TDE for source localisation estimation and here the MUSIC algorithm [1] is used. Due to the use of highly directional sensors, the AVS provides many advantages over other microphone arrays for DOA estimation. In particular, the secondary reflections in reverberant conditions are minimised due to two features of the array, (a) the co-location of the sensors, and (b) the directionality of the sensors. There are post-processing techniques for improving the localisation accuracy for spaced microphone arrays [8, 9, 10]. However, in this work, the focus is on investigating the advantages that can be drawn from the AVS without such post-processing techniques. The motivation is to minimise additional computational complexity for use in real time applications such as speech teleconferencing. To the best of the author's knowledge this is the first time a single collocated microphone array is used for DOA estimation of speech sources in reverberant conditions and for a moving target.

The paper is organised as follows section 2 presents the method used for estimating the DOA for an AVS, section 3 presents the experimental setup and the results for DOA for monotone stationary sources. In section 4 results of stationary and moving speech sources are presented and in section 5 results and conclusions are presented.

II. SOURCE LOCALIZATION USING AVS

A. AVS Array

The output of the AVS consists of two components: an acoustic particle velocity and acoustic pressure component. This can be expressed in vector form as:

$$\mathbf{y}(t) = [p(t), v_x(t), v_y(t), v_z(t)]^T \quad (1)$$

978-1-4244-7907-8/10/$26.00 ©2010 IEEE

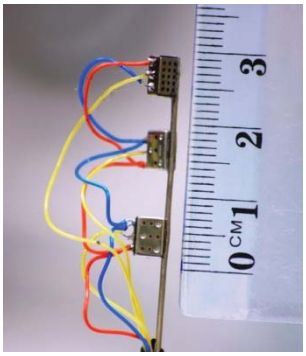

Fig 1: The AVS of [3] used in this work

Where $p(t)$ represents the acoustic pressure component and $v_x(t)$, $v_y(t)$, $v_z(t)$ represents that the velocity gradient of the x, y and z axis. The relationship between the acoustic pressure and the particle velocity is given by [11]:

$$v(r,t) = f\left(p(r,t)\right)\mathbf{u} \tag{2}$$

Where $v(r, t) = [v_x(t), v_y(t), v_z(t)]$ represents the acoustic particle velocity vector and f is a function of the acoustic pressure gradient and where:

$$\mathbf{u} = \left[\cos\theta\sin\phi \quad \sin\theta\cos\phi \quad \sin\phi\right]^T \tag{3}$$

is the source bearing vector with θ representing the azimuth and ϕ the elevation [11].

B. The SoundField Microphone

Unlike the AVS the SoundField microphone uses 4 cardioid capsules arranged in a regular tetrahedron configuration. The out puts from the capsules are added and subtracted to get the B format output which is similar to the output of the AVS. The four components of the B format are formed as follows.

The capsule array has four capsules termed left front (LF), right back (RB), right front (RF) and left back (LB). To form a figure of eight in the *x* direction, the left front is subtracted from the right back to form a figure of eight in the horizontal with axis along the line left front and right back. The right front and left back are subtracted to form a figure of eight in the horizontal line along that line. The two diagonal figures of eights are subtracted to form the X component and similarly the Y and Z components are formed [12]. Similar to (1) for the AVS, this result in a set of pressure and directional recordings, which are formed as follows:

$$W = \text{LF} + \text{RB} + \text{RF} + \text{LB} \tag{4}$$
$$X = \text{LF} - \text{RB} + \text{RF} - \text{LB} \tag{5}$$
$$Y = \text{LF} - \text{RB} - \text{RF} + \text{LB} \tag{6}$$
$$Z = \text{LF} + \text{RB} - \text{RF} - \text{LB} \tag{7}$$

Equations (4)-(7) represent the W, X, Y and Z components of a standard first order Ambisonic B-format recording as produced by the SoundField microphone [12]. Similar to an AVS, the equations (2) and (3) will hold for the SoundField microphone, where microphones X, Y and Z represent directional components of the pressure gradients and W is just an amplitude scaled pressure recording [13].

C. DOA Estimation using Multiple Signal Classification (MUSIC)

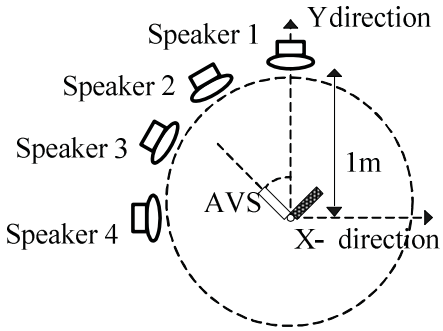

Fig 2: Experimental Setup,

This algorithm uses only the velocity components of the AVS that is the X, Y and Z components and first estimates the covariance matrix of the velocity components. This can be expressed as follows [11]:

$$R = \frac{1}{N}\sum_{t=1}^{N}\text{Re}\left\{\mathbf{y}_\mathbf{v}(t)\mathbf{y}_v^*(t)\right\} \tag{12}$$

From (12), $\bar{\mathbf{u}}$, defined as the unit eigenvector of R associated with the largest eigenvalues of R, can be used to estimate the source bearing vector $\mathbf{u}$ of (3) following the rules outlined in [11]. Here, $\mathbf{y}_v(t)$ is vector containing the X, Y and Z components for AVS and the SoundField microphone. A significantly more reliable and efficient method for finding the DOA estimate is the MUSIC algorithm of Schmidt [1]. The MUSIC algorithm allows for the estimation of the DOA using the eigenvalues and eigenvectors of the covariance matrix formed from the recorded signals [1]. The MUSIC algorithm can be expressed as:

$$\theta = \min_{\theta_i}\left[P(\theta_i) = \frac{1}{\sum\left|\mathbf{V}_i^H\mathbf{h}(\theta_i)\right|^2}\right] \tag{13}$$

where $\mathbf{V}_i^H$ is the smallest eigenvector of the covariance matrix R of the pressure and velocity components of the AVS and $\mathbf{h}(\theta_i)$ is the steering vector for the AVS and where $\theta_t \in (-\pi, \pi)$. For sources with an elevation of 0 relative to the AVS (assumed in this work), the steering vector [4] can be described as a function of the azimuth as:

$$\mathbf{h}(\theta_i) = \left[\cos(\theta_i) \quad \sin(\theta_i) \quad 1\right] \tag{14}$$

which is formed from the x and y components of (3) with $\phi = 0$ and where 1 represents the omni-directional microphone. The peaks of $P(\theta_i)$ represent the DOA estimate for that source. In this work the MUSIC algorithm is used. Where only the velocity components of the AVS output are used in the MUSIC algorithm for estimating the DOA, this reduces the size of the covariance matrix calculated hence reducing the computational complexity of the DOA estimation.

III. Localisation Experiments

A. Experimental Setup

Recordings were made in reverberant room with a RT_{60} of 30ms and with considerable background noise of computer servers and air-conditioning at 53.1dBA. For testing, the experimental setup of Fig. 2 was used, where the AVS was mounted on a custom built rotating platform (to allow positioning of the microphones relative to the source) and a self powered

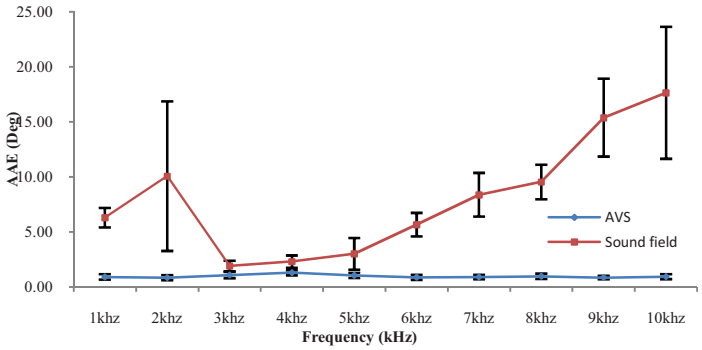

Fig 3: AAE for DOA estimates for AVS and Soundfield

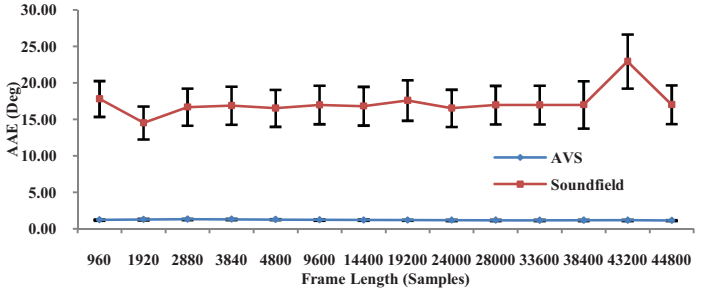

Fig 4: AAE for DOA estimates for different frame sizes

loudspeaker (Genelec 8020A) was place in front of the AVS at a distance of 1m with an elevation of 0^0. A series of monotone signals each 2 seconds long and of equal energy were played with frequencies ranging from 1 kHz to 10 kHz. For speech 5 male and 5 female sentences from the IEEE speech corpus [14], each approximately 2.5 s long with different speeds were played. Recordings were made at 5^0 intervals and signals were sampled at 48 kHz.

To simulate moving targets, 3 additional loudspeakers were used as shown in Fig. 2. The average speech of walking for a human being is 1.33m/s [15]. This means on average in a circular path with a radius of 1m a man walking at this average speed would take 0.13s to walk 10^0. The speech sentences were sliced into four parts each part 0.066s long for fast moving, 0.13 s for normal walking speed and 0.3 s for slow walking paces and the speakers are separated by 30^0. Each part of the sentence is played on one loudspeaker in order and between each part a silence of approximately 0.2s for fast moving, 0.4s for average walking speed and 0.8s for slow walking is introduced. Hence, the experimental setup simulates a source moving over 4 sectors, each covering 10^0.

The results present in this work are for average angular error which is the error between the actual angle and the angle obtained from the DOA estimate, which is calculated as follows:

$$AAE = \frac{1}{N}\sum_{n=1}^{N}\left|\theta_{n,m} - \theta_{n,a}\right| \qquad (15)$$

where N is number of sources (tones) and $\theta_{n,m}$ and $\theta_{n,a}$ are the measured, m, and actual, a, DOAs, respectively, for source n. The results presented in following sections are for confidence intervals of 95 %.

B. Monotone Stationary Sources

Fig. 3 shows the results for AAE for monotone signals over a rotation of 90^0 in azimuth at 5^0 intervals for the AVS and

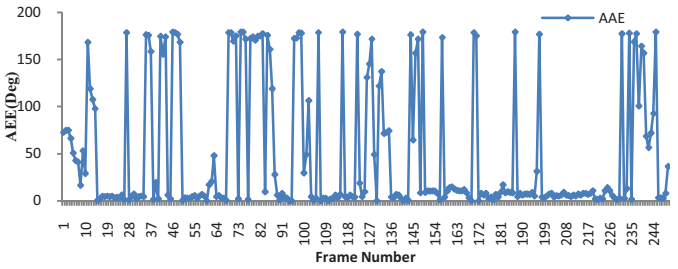

Fig 5: AAE for DOA estimate for each frame of a speech sentence

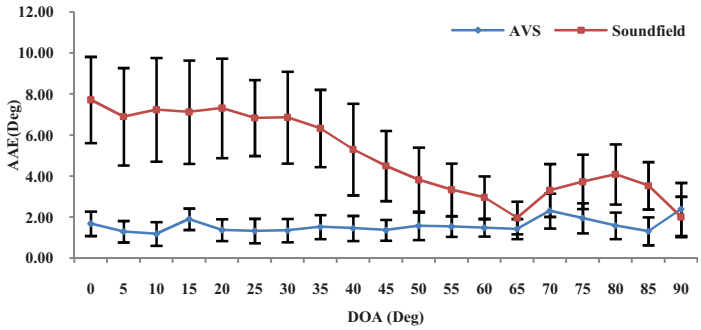

Fig 6: AAE for DOA estimates for Speech Sources

SoundField microphones. The DOA estimates obtained are for the average of the all the frames of the recorded signal. The frame length used is 20 ms or 960 samples. The results show that the AVS has an average error of 0.98^0 and the average error for the sound field is 8.2^0.

The SoundField uses all the cardioid capsules to generate the directional components, hence capture reflections from all the directions. These reflections are then included as error in the formation of the X, Y and Z components. The other important factor that affects the results is the influence of the protective netting of the sound field as these would diffract and reflect the sound signals. In [4] it was found that for the AVS the mount and the positioning of the microphone capsules contributed to errors in DOA estimates. In addition for an omni-directional microphone which has no directional bearing on the output, there is relationship between the aperture of the capsule and the frequency of the signals that is if the wave length of the signal is smaller than the aperture then the omni-directional microphone will start to display directional characteristics [13]. As seen from the results the SoundField produces larger errors at higher frequencies especially above 8 kHz which is the frequency at which most omni-directional capsules start to exhibit the directional characteristics [13]. This change in the polar pattern may be a reason for the increased inaccuracy of the DOA estimate from the SoundField microphone.

Results in Fig. 4 are for the varied frame lengths from 960 samples (20ms) to 48000 samples (1 s), this is done to find out if it is possible to estimate the DOA from a single frame and if so what is the smallest frame length the will give an accurate results. The monotone signals are of equal energy for the entire duration, hence the DOA estimates from single frame should be approximately the same as that of the average. For a signal which has time varying energy, like speech, the DOA estimates from each frame may be different and especially if the source is moving. Hence it is crucial to find the smallest frame length at

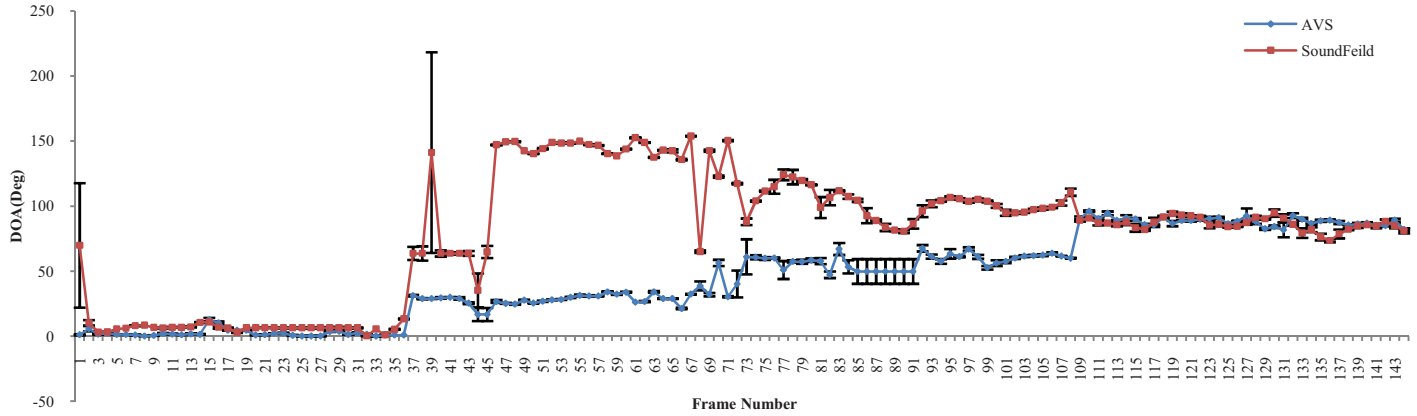

Fig 7: DOA estimate for slow moving speech source

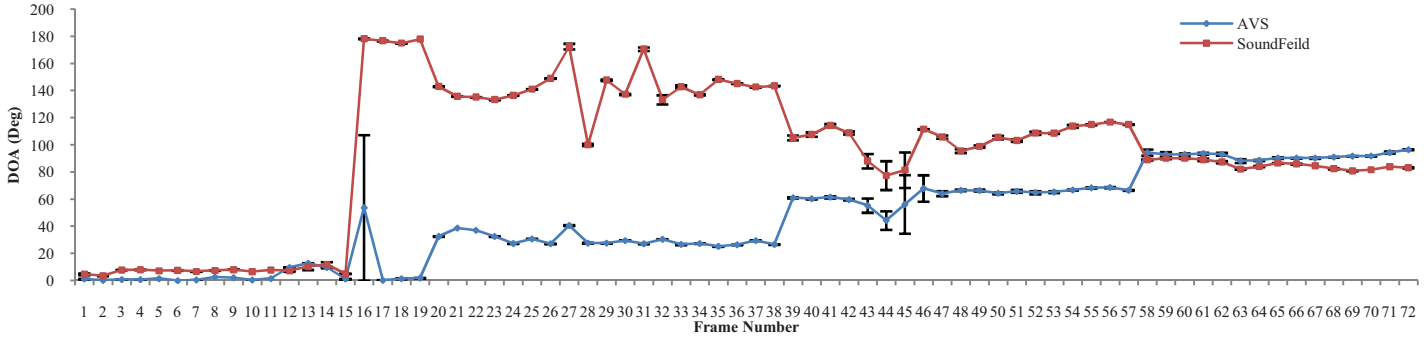

Fig 8: DOA estimate for Normal moving speech source

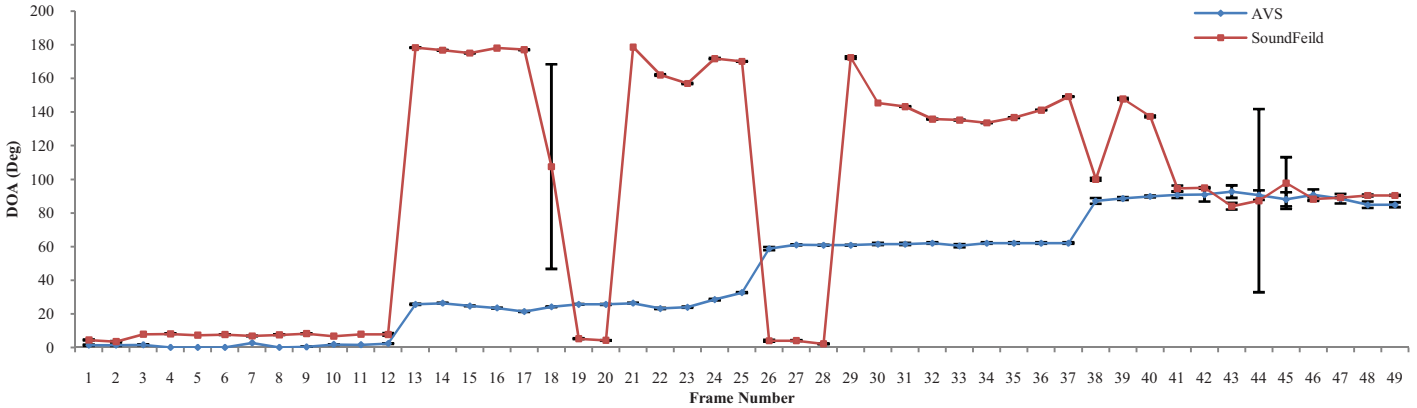

Fig 9: DOA for Fast Moving speaker source

which an effective DOA estimate can be obtained for a single frame.

The results show that there is no effect from the frame length on the outcome of the DOA estimate, but the AAE for a single frame is higher than that for all the frames averaged. The AAE for a single frame is 1.21^0 for the AVS and 17.23^0 for the sound field. These results confirm that with a single frame of 20ms it is possible to obtain an accurate DOA estimate.

IV. DOA ESTIMATION FOR SPEECH

A. Stationary Speech Sources

Unlike monotone signals speech has different characteristics. The energy of the speech signal varies over time, there are voiced, unvoiced and silence in the sentence which should be considered. From Section III it has been established that frame lengths of 20ms are enough to get an accurate DOA estimate. The results presented in Fig. 5 are for all the frames of a speech sentence at a 0^0 azimuth to the microphone with a frame length of 960 samples or 20ms. The results show that all the regions of the speech which are unvoiced or stops produce errors and the AAE is 49^0. This is expected as these regions are similar to that of no speech. In order to fix this error, a modified version of the VAD based on ITU-T G.729B [16] is introduced to the algorithm. The VAD flags any frame that is unvoiced or if it is a stop. Since the frame length is 20ms it is assumed there is no significant change in the position of the speaker in 20ms and that frame is given the DOA estimate of the previous frame.

Results in Fig. 6 are for the DOA estimation for speech sources with VAD implemented in the algorithm. The results show that with the VAD in place the AAE for the AVS is 1.58^0 and for Soundfield the AAE is 4.99^0. These results show that there is a significant influence from the unvoiced and stop sections of the speech on the DOA estimate.

B. Moving Speech Sources

The time taken for person moving through a 10^0 arc is larger than the frame length required for producing an accurate DOA estimate. The time taken for an average person to walk an arc of 10^0 at a distance of 1m from the microphone is 0.13sec, which is 6.5 frames at a 48 kHz sampling rate and frame sizes of 960 samples. But because the speech has unvoiced sections and stops, a single frame is insufficient to produce an accurate DOA estimation as the frame may be unvoiced or a silence. Hence, the length of the speech segments in each speaker is at least 6 frames long and the time taken for the speech segment to move from one loudspeaker to the next is introduced as mentioned in section III *A*. The frame length used in calculating the DOA estimate for the fast moving source is 480 samples or 10ms as it was found that to get sufficient number of frames for fast moving source frame size of 960 only had 3 frames hence to get at least 6 frames the frame size is reduced.

The results presented in Fig. 7 are for those of a source moving at slow walking speed of 0.665m/s, the speech on each loudspeaker is stationary for 0.03s. The results for the AAE for each stop section for the AVS and sound field are shown in Table 1.

The results presented in Fig. 8 are for those of the source moving at normal walking speed of 1.3m/s, the speech on each loudspeaker is stationary for 0.13s. The results for the AAE for each stop section are shown in table 1. Similar results for the fast walking speaker is shown in Fig. 9 and the AAE is shown table 1, the stop section on each loudspeaker is 0.0066s.

	0	*30*	*60*	*90*
AVS– Fast	1.2	5.6	1.5	5.7
SF – Fast	7.8	112.7	81.9	10.5
AVS – Nor	3.9	5.1	5.7	3.8
SF – Nor	69.9	99.7	46.1	6.2
AVS - Slo	3.7	4.5	4.0	11.1
SF - Slo	16.3	93.1	42.5	10.6

Table 1: AAE of moving source for AVS and SoundField

V. CONCLUSION

The results obtained for the DOA estimation with AVS and SoundField microphone shows that AVS is capable of providing DOA estimates for stationary and moving speech sources with AAE's error's of 1.58^0 while for soundfield the AAE is at 4.99^0. The accuracy of AVS is reduced for moving speech sources and AAE increased from 1.58^0 to an average of 4.6^0 and similarly the error for the SoundField microphone also increased for moving sources from 4.99^0 to 49.76^0. Although the error for moving source has increased for the AVS, the error is less than 5^0. Further, the results show that AVS is capable of making accurate DOA estimates with frame sizes of 20 ms for moving sources.

The results show that the AVS has the ability to give highly accurate DOA estimates in reverberant conditions for stationary and moving speech sources. This result is very important as to track a moving source, geometrically spaced microphones are normally used. The AVS compared to a SoundField microphone has better performance in terms of DOA estimation. Future work will investigate the application of this work to the enhancement of moving speech sources where accurate DOA estimation is critical.

Acknowledgement

This project was partially supported by the Australian Research Council Grant DP0772004.

REFERENCES

[1] R. O. Schmidt, "Multiple Emitter and Signal Parameter Estimation," *Proceedings of RADC Spectral Estimation Workshop*, 243-258, October 1979.

[2] M., S., Brandstein, and D., B., Ward, "Microphone Arrays: Signal Processing Techniques", Berlin: Springer-Verlag, 2001.

[3] T., S., Brades ahd R., H., Benson "Sound source Imaging of low flying airbone targets with an acoustic cemera array", *Appl. Acoust.,* 68, 752-765, 2007.

[4] M. Shujau, C.H. Ritz, I.S. Burnett, "Designing Acoustic Vector Sensors for localization of sound sources in air", *EUSIPCO 2009*, UK, August 2009.

[5] S., Mohan, M.,E., Lockwood, M., L., Karmer, and D., L., Jones, "Localization of Multiple Acoustic Sources With Small Arrays Using a Coherence Test", *J. Acoust. Soc. Am*, 123, 2136-2147, April 2008.

[6] M., Kawanishi, R., Maruta, N., Ikoma, H., Kawano., H., Maeda, " Sound Traget Tracking in 3D using Particle Filter with 4 Microphones", *SICE Annual Conference 2007*, Japan, September 2007.

[7] H., Atmoko, D., C., Tan, G., Y., Tian, and B., Fazenda, " Acurrate sound source localization in a reverberant enviroment using multiple acoustic sensors", *Measurement Science and Technology*,19 (2008), January 2008.

[8] N., Roman, and D., Wang, "Binural Tracking of Multiple Moving Sources", *IEEE Trans. On Audio, Speech and Language Processing*, Vol. 16, No. 4, May 2008.

[9] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberent enviroments", Proc. ICASSP, Vol. 5, pp. 3021-3024, 2001.

[10] D., B., Ward, E., A., Lehmann and R., C., Williamson, ",Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Enviroment", *IEEE Transactions on Speech and Audio Processing*, , vol.11, no.6, pp. 826- 836, Nov. 2003.

[11] A., Nehorai and E., Paldi, "Acoustic Vector Sensor Array Processing", *IEEE Transactions on Signal Processing*, Vol. 42, No.9, 2481-2491, September 1994.

[12] User Manual for ST 250, Sound field reserch Ltd, West Yorkshire, England, Issue 1.5.

[13] J. Eargle, "The microphone Book", Boston: Focal Press, 2001.

[14] IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio and Electro-acoustics*, AU-17(3), 225-246.

[15] J. M., Burnfield, and C., M., Powers, "Normal and Pathologic Gait, in Orthopaedic Physical Therapy Secrets", Hanley & Belfus; 2 edition, June, 2006.

[16] Voice Activity Detector implemented by Prof. Peter Kabal, available online at, http://www-mmsp.ece.mcgill.ca/Courses/2007-2008/ECSE412B/Project/MATLAB/VAD.m